DOCUMENT RESUME

ED 178 585 TM 009 771

AUTHOR Kreines, David C.; Mead, Ronald J.
TITLE Equating Tests With the Rasch Model.
PUB DATE 11 Apr 79
NOTE 17p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 9-11, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Complexity Level; *Equated Scores; Goodness of Fit; *Item Analysis; Mathematical Models; Raw Scores; Statistical Analysis; Test Construction; Test Interpretation; Test Items
IDENTIFIERS Rasch Model; *Rasch Scaled Scores; *Test Equivalence

ABSTRACT

An explanation is given of what is meant by "sample-free" item calibration and by "item-free" person measurement as these terms are applied to the one-parameter logistic test theory model of Georg Rasch. When the difficulty of an item is calibrated separately for two different samples the results may differ; but, according the the model, the estimated item difficulty is invariant except for an arbitrary constant. Test equating with the Rasch model is the process of determining this constant. Two methods of test equating are described and illustrated by simple numerical examples and by equivalent graphic figures. Common-person equating is done when the same group of persons takes two different tests. Since the persons are assumed to have the same ability regardless of which test they take, the scores on the more difficult test may have a constant added to make them equal to equivalent scores on the easier test. Common-item equating assumes that a given item has only one difficulty. When two tests contain some items, common to both, then common item equating is appropriate. Any difference in the difficulty of the other items in the two tests may be adjusted by calibrating through the common items. Tests of the goodness of the fit of the data to the model are discussed. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED178585

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

David C. Kreines

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

EQUATING TESTS WITH THE RASCH MODEL

DAVID C. KREINES, National Board of Medical Examiners
RONALD J. MEAD, University of Chicago

APRIL 11, 1979

AERA/NCME JOINT SESSION

NUMBER 22.20

IM009 771

EQUATING TESTS WITH THE RASCH MODEL

DAVID C. KREINES, National Board of Medical Examiners
RONALD J. MEAD, University of Chicago

INTRODUCTION

The measurement philosophy of Georg Rasch tells us that every item has a difficulty that is inherent in it regardless of the ability level of the calibration sample. Analogously, every person has an ability that is independent of any set of items that might sometime be used to measure it. And, according to Rasch, it is possible to estimate these parameters with no more assumptions than those made in standard practice. This "sample-free" item calibration and "item-free" person measurement is very attractive, but the words have sometimes led to confusion.

We could, for example, administer a test, calibrate the items and learn that the "sample-free" item difficulty for item one is 1.5 logits. But when we use this same item in another form with another sample, we find its difficulty to be different, say -0.5 logits. In what sense then can we say that this is invariant?

Mathematically, the answer is simple: The estimated item difficulty is invariant except for an arbitrary constant. Test equating with the Rasch model is the process of determining this constant.

The model, however, is not magic. Two tests which have nothing in common cannot be connected, but we do have considerable freedom in what the common ground can be. It may be that the tests were given to the same people, or it may be that they contain some of the same items. The illustration just mentioned with two forms linked by a single item represents the simplest possible situation.

In order to understand how to equate forms, we need to be clear about exactly what the 1.5 and -0.5 actually mean. Because the estimation process requires one arbitrary constraint, most calibration programs (i.e. BICAL)

constrain the difficulties to average zero, solely for numeric reasons. What this means for us is that to say an item has a difficulty of 1.5 logits in form A is to say that it is 1.5 logits more difficult than the average item in that form. The -0.5 means that this same item is half a logit less difficult than the average item in form B.

The equation of the two forms is then simple (see Figure 1). If we leave the origin at the center of form A, we need only establish the distance to the center of form B. Moving from 0.0 to the common item is 1.5 logits (i.e., its difficulty on form one). From the common item to the center of the second form is plus 0.5 logits (plus because we are moving from the item to the center). Hence, the distance from the center of form A to the center of form B is

$$1.5 + 0.5 = 2.0 \text{ logits.}$$

To complete the equation of form B to form A, we need only add this constant to every item difficulty from form B and to every ability obtained from form B. Therefore, the common item's new form B difficulty is

$$(-0.5 + 2.0) = 1.5,$$

just as it is in form A. The average item difficulty is

$$(0.0 + 2.0) = 2.0.$$

The relationship between the center of the form and the common item is unchanged:

$$(1.5 - 2.0) = -0.5.$$

In the same way an ability estimated to be 1.0 based on form B is equivalent to an ability of 3.0 from form A.

As can be seen from Figure 1, the invariance in this scale is in the relationship between points. Any numbers we like can be attached to the points as long as the differences are not changed. Common person equating and common item equating lead to different numbers for the points

because different origins are chosen, but the principle involved is the one we have been discussing: A common reference point is specified and the distances from there to any interesting point is determined.[1]

COMMON PERSON EQUATING

When the same group of persons takes two different tests, which we may refer to as Form A and Form B, the two tests may be equated by a technique referred to as "common person equating". As a starting point, we assume that any person v has one "god-given" ability β_v for which we obtain two estimates. If the person's ability estimate resulting from his raw score on form A is different than the ability estimate resulting from his raw score on form B, then this apparent difference must be due to some difference between the two tests, namely their relative difficulties. For example, if an ability estimate of 1.0 logit is obtained for person v on test A, while test B yields an ability estimate of .5 for the same person, then we know that form A is 1.0 logit below the person while form B is 0.5 logits below. The person is the reference point in both cases. Therefore, we can calculate the distance from form A to form B as

$$(1.0 - 0.5) = 0.5$$

or, in other words, form B is 0.5 logits more difficult than form A. The equation is completed by adding half a logit to each form B difficulty (and ability). Of course, we have more confidence in the result if we use more than one person.

As an example, we may find that test A estimates a mean ability of 1.5 logits for the calibration while test B estimates a mean ability of

---

[1] It is often convenient to refer to the test being equated (or adjusted) as the "current test", and the test serving as the common scale as the "reference test".

1.0 logit (see Figure 2). Keep in mind that the calibration process forces the mean of the item difficulties to be 0, and that these estimated mean abilities are with respect to this artifically determined origin. Since test A and test B are estimating the same ability, we can "slide" the logit scale so that the two estimated abilities line up on the scale. One way to accomplish this is to redefine the origin to be the mean ability, as shown in Figure 3. Since the relationship between the mean ability and mean item difficulty for a test must remain fixed, we must also shift the location of the mean item difficulties on the same scale. Since it is actually the item difficulties that we are interested in, one way this may be accomplished is by subtracting the mean ability obtained as a result of calibrating the test from each of the item difficulties, which is equivalent to subtracting the mean ability from the mean item difficulty, which is always set to 0. It can be seen then that the difference between the difficulties of the two test forms is simply the difference between the two mean ability estimates. From Figure 3, it can be seen that test B is .5 logits more difficult than test A, and that in order to consider the two tests to be equivalent, either .5 logits must be added to each item difficulty on test A, or .5 logits must be subtracted from each item difficulty on test B.

COMMON ITEM EQUATING

While common person equating rests on the assumption that a particular person has only one true ability which may be estimated by more than one test, common item equating assumes that a given item has only one difficulty, which may be estimated by different groups of persons to whom that item is administered. When two tests contain some items,

common to both, then common item equating is appropriate. The subset of items which appears in both tests is referred to as the "link" between the tests. Because these items have only one true difficulty, then any difference observed in the difficulties of these items as a result of the calibration of the two tests must be due to the relative difficulty or easiness of the other non-link items in the respective test. This fact may then be used as the basis of a adjustment in the scale in order to correct for this perceived difference.

Figure 4 shows the mean item difficulty for test A, which is administered to group 1, and the mean item difficulty for test B, which is administered to group 2. Also shown is the mean difficulty of the link items AB which are included in both tests. Keep in mind once again that the origin for the scale is defined by the mean item difficulty for each test, and that the mean difficulty of the link items in each test is set with respect to that arbitrary origin. Since we assume that the mean difficulty of the linking items is really the same, then we may use the observed mean difficulties of the links to establish a new origin, which will then make the differences in difficulty between the two tests apparent. Such a shift is accomplished by subtracting the mean difficulty of the link items from the difficulty of each item on the test, and the result of this change in the origin is shown in Figure 5. The distance on the scale between the mean item difficulty of the whole test and the mean item difficulty of the link items remains the same, but it is now apparent that test A is .5 logits more difficult than test B. In order to make the two tests comparable, then, we must either add .5 logits to each item difficulty on test B, or subtract .5 logits from each item difficulty on test A.

Naturally, this process is much more complex when you are faced with

a test with more than a few linking items. To help speed up this process, we have developed a computer program called LINK, which uses as its raw data two sets of calibrated item difficulties and their corresponding item names. The program uses the item names to identify those items which belong to both tests, which we call the link. It then calculates the difference between each pair of item difficulties. The average difference is then the distance between the origins of the two tests.

TESTS OF FIT

As we have seen, we can determine the distance between two tests with one piece of information: Either the ability of a common person as measured by the two tests or the calibration of a common item within each test. If we have replication, either many people or several items, we gain confidence in our estimate of the difference. But at least as important, we gain the opportunity for assessing the "validity" of the link, that is, a statistical test of fit.

The tests of fit are extraordinarily simple. After equating the two tests, each common item has two difficulties which should be the same number. Plotting the form A difficulties against the form B difficulties, after equating, gives a line of slope one through the origin. Confidence bounds can be drawn around the link and the items which fall outside the bounds are readily identifiable.

Any number of statistics can be computed to describe this line and how closely the items are to it. None work any better than the simple graphical approach. A Z-statistic based on the difference between the equated item difficulties divided by the standard error of the difference is an obvious choice.

$$Z = \frac{(d_{Ai} - d_{Bi}) - T_{AB}}{SE_T}$$

However, since the standard errors are smallest for central items, this statistic is less tolerant of discrepancies for these items than for extreme ones although the impact on the estimate of the mean is the same.

Correlations and regression slopes are less useful. The correlation will always be near one and takes no account of the fact that the dispersion in the two sets of difficulties must be the same. The regression slope ignores the symmetry in the error distributions which result in the slope appearing to be different than one when all items actually do fit.

Although we haven't yet done it, a censoring procedure based on robust statistics would probably provide an effective computerized alternative to looking at a plot. Our motivation on this topic is low, even though there often is more noise in the system than the model expects, since it tends to cancel itself out and the mean difference, which is what we are actually after, is quite stable.

The fit analysis is interesting from another point of view. It has been our experience that when an item is clearly off the line, we were mistaken in considering it to be the same item. These always have explanations like the item was miskeyed in one form, the item referred to a chart that was poorly reproduced in one, or a subtle change was made in one of the distractors that had a dramatic impact on the item's difficulty. This sort of quality control has been found very valuable and often very enlightening in terms of what affects an item's difficulty.

COMPARING RAW SCORES

Now we have seen how to adjust the individual item difficulties for a given test in order to make it comparable with another test B. But what is the practical significance of this? What we often need to know

is what raw score on test A is equivalent to another raw score on test B. The Rasch model provides a seemingly complex but actually simple method of generating this score equivalence table.

Given a set of calibrated item difficulties, it is possible to generate, for each raw score, the corresponding Rasch ability. By using the calibrated difficulties on test A as input to this process, we can create a raw score ability table. Likewise, by using the difficulties from test B, to which an adjustment has been made to align them with test A, we may create another raw score ability table. Table 1 is an example of such a table for two six-item tests. Since we have adjusted the difficulties for test B to place it on the same scale as test A, then it must follow that any particilar ability estimate on either test is an estimate of the same true ability. If, therefore, two abilities are the same, then the raw scores corresponding to those abilities must also be equivalent. As an example, in Table 1, the ability of .5 logits corresponds to a raw score of 4 on test A, and a raw score of 5 on test B. Therefore, we know that these two scores are equivalent. A person must get five items correct on test B to perform at the same level as a person getting four items correct on test A. This same procedure is followed for each ability estimate on the test, and the result is an equating table as shown in Table 2.

Once again, it can be seen that this would be a rather complex and time consuming process to do by hand, but with the aid of the computer it becomes relatively simple. We have developed a computer program called EQUATE which performs the equating based on a "shift constant" provided by LINK, and produces the score equivalence tables.

In summary, the Rasch approach to test equating is very attractive because of the unique measurement properties of the model. The model is

first very economical in its data requirements and second it has sufficient checks to warn us of possible dangers. The remaining papers will provide discussions of how effective these were in real situations.

REFERENCES

Rentz, R.R. and Bashaw, W.L. Equating Reading Tests with The Rasch Model, Volume 1 Final Report. Volume II Technical Reference Tables. Athens, GA: University of Georgia, Educational Research Laboratory, 1975. (ERIC Document Reproduction Nos. ED 127 330 through ED 127 331).

Rentz, R.R. and Bashaw, W.L. The National Reference Scale for Reading: An Application of the Rasch Model. Journal of Educational Measurement, 1977, 14 (2), 161-179.

Wright, B.D. Sample-Free Test Calibration and Person Measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service, 1968, 85-101.

Wright, B.D. Solving Measurement Problems with the Rasch Model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B.D. and Mead, R.J., R. J. BICAL: Calibrating Items and Scales with the Rasch Model. Research Memorandum No. 23a; Statistical Laboratory, Department of Education, University of Chicago, 1978.

EQUATING TESTS WITH THE RASCH MODEL

DAVID C. KREINES, National Board or Medical Examiners
RONALD J. MEAD, University of Chicago

FIGURE 1

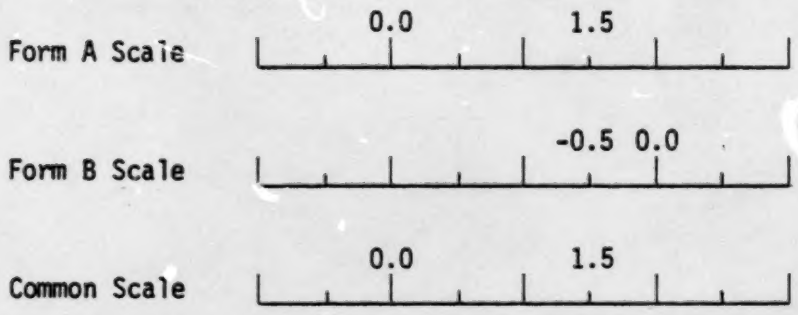

FIGURE 2

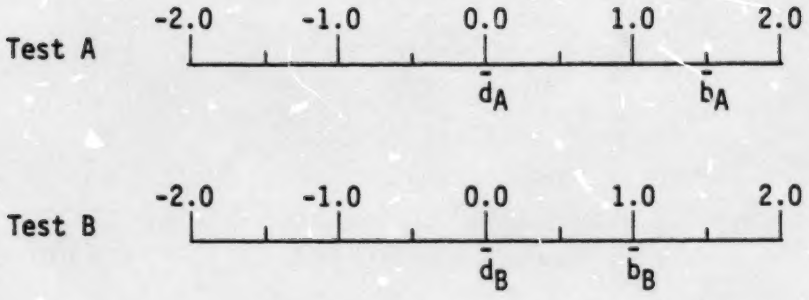

FIGURE 3

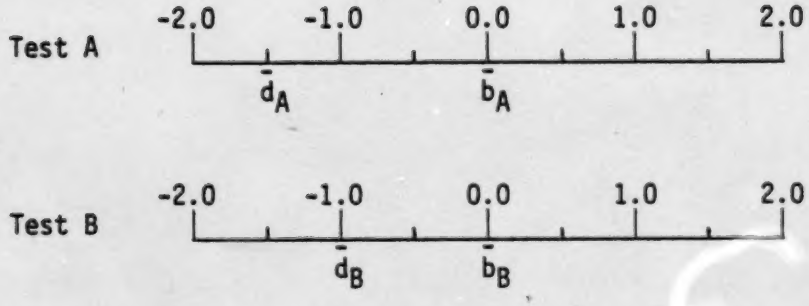

FIGURE 4

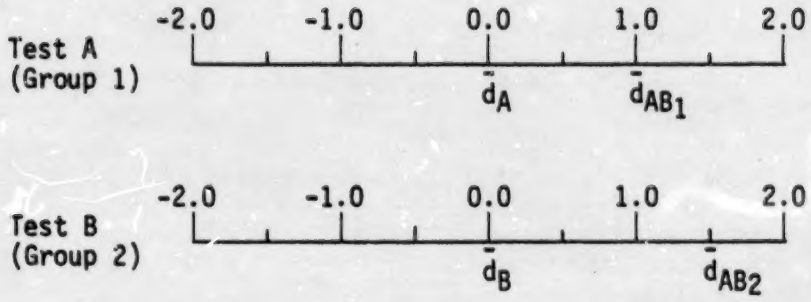

FIGURE 5

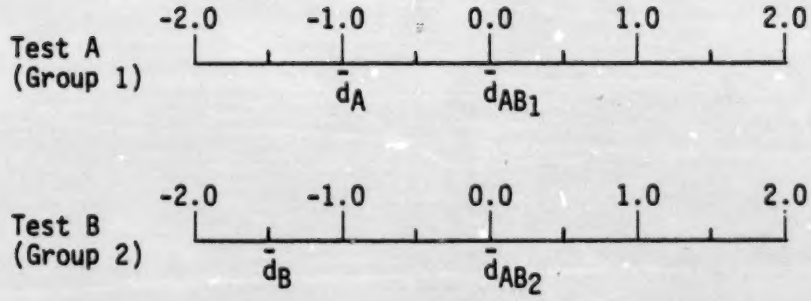

TABLE 1

Form A		Form B	
Raw Score	Logit Ability	Raw Score	Logit Ability
1	-1.0	1	-1.5
2	-0.5	2	-1.0
3	0.0	3	-0.5
4	0.5	4	0.0
5	1.0	5	0.5

TABLE 2

Form A Raw Score	Form B Raw Score
1	2
2	3
3	4
4	5
5	-